UNIVERSITY OF ALBERTA

DIFFERENCE IN GENDER DIFFERENTIAL ITEM FUNCTIONING PATTERNS ACROSS ITEM FORMAT AND SUBJECT AREA ON DIPLOMA EXAMINATIONS AFTER CHANGE IN ADMINISTRATION PROCEDURE

by

YANQI FENG

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Education

In

Measurement, Evaluation and Cognition

Department of Educational Psychology

Edmonton, Alberta

Fall, 2008

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-46998-9
Our file Notre référence
ISBN: 978-0-494-46998-9

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

The primary purpose was to determine whether the new administration procedure, where the constructed response items and the multiple-choice and numerical response items were administered on two occasions, had any differential impact on the gender differential item functioning (DIF) patterns across content areas and item formats on the Diploma Examinations in Alberta. A secondary purpose was to examine potential sources for gender differences using a common sample of examinees across examinations and item formats. Data from three different Alberta Education Diploma Examinations, namely Social Studies 30, Biology 30, and Pure Mathematics 30, were analyzed. A mixture of dichotomous and polytomous items were used in each of the examination studied. The samples included students who wrote all three examinations in June 2002, when the Diploma Examinations were administered using the old administration procedure, and the students who wrote all three examinations in June 2005, when the examinations were administered using the new administration procedure. Poly-SIBTEST was the DIF detection method used. The results revealed that the prevalence of DIF and the patterns of DIF within content areas were similar across the four years examined. The change in administration schedule did not lead to a change in the prevalence and patterns of gender DIF. However, an item format effect where females performed better than males of the same ability on polytomous items was observed across subjects for the fours years studied. The findings of the study provide useful insights for policy makers to evaluate the new administration procedure as well as the diploma examination program.

Acknowledgement

I would like to take this opportunity to thank a few people who have helped me achieve this goal. First of all, I would like to thank my supervisor, Dr. W. Todd Rogers. Dr. Rogers has provided detailed comments on several earlier versions of this thesis, helping me clarify ideas, organize thoughts and accurately present research findings. I have learnt a lot through working on my thesis under his supervision, not only knowledge and skills to conduct research in the area of psychometrics, but also a positive attitude, patience and enthusiasm towards research in this field. I could never thank him enough for all the guidance, support and encouragement I received throughout my master study.

Secondly, I would like to thank Dr. Mark J. Gierl who introduced me to various practical issues facing psychometricians in the educational testing area, including IRT, DIF, Equating, computer adaptive testing. I would also like thank him and Dr. Carolyn Ross for serving as my committee members and providing valuable comments and suggestions for my thesis.

I am very grateful to many fellow students who studied at the same time with me in CRAME, particularly, Ying Cui, Ling Peng, Xuan Tan, and Jiawen Zhou, whose company has made my master study a much more fun and pleasant experience.

Lastly, I would like to thank my family members, particularly my parents, for the unconditional love and support they have given me throughout all my endeavours.

Dedication

To my father

Table of Contents

CHAPTER 1: INTRODUCTION **1**

Purpose of the Study 5
Definition of Terms 5
Organization of the Thesis 6

CHAPTER 2: LITERATURE REVIEW **8**

Gender DIF across Content Areas 8
Gender DIF across Item Formats 10
Overview of the Poly-SIB procedures 11
Change of Administration Procedure of the Diploma Examinations 15

CHAPTER 3: METHOD **17**

Diploma Examinations 17
Subject Areas 18
Samples 20
Procedure 21

CHAPTER 4: RESULTS **22**

Differences between Males and Females on Sets of Items 22
Social Studies *23*
Biology *24*
Pure Mathematics *24*
Prevalence of DIF across Subject Area and Item Format 25
Social Studies *25*
Biology *26*
Pure Mathematics *26*
Gender DIF Patterns across subject area and item format 27
Social Studies *27*
Biology *30*
Pure Mathematics *34*

CHAPTER 5: SUMMARY AND CONCLUSIONS **38**

Summary of Research Questions and Method 38
Summary of Findings 40
Prevalence of DIF *40*
Potential Sources of DIF *41*
Limitations of the Study 43
Conclusions 45
Implications for Practice 45
Recommendations for Future Research 45

REFERENCES **49**

APPENDIX A **55**

List of Tables

TABLE 1.STRUCTURE OF EXAMINATIONS 18

TABLE 2. DESCRIPTIVIE STATISTICS FOR SOCIAL STUDIES EXAMINATION ACROSS YEARS..... 23

TABLE 3. DESCRIPTIVIE STATISTICS FOR BIOLOGY EXAMINATION ACROSS YEARS 24

TABLE 4. DESCRIPTIVIE STATISTICS FOR PURE MATHEMATICS EXAMINATION ACROSS YEARS..... 25

TABLE 5. DIF ITEMS ACROSS EXAMINATIONS AND ITEM FORMATS..... 26

TABLE 6. SOCIAL STUDIES DIF ITEMS BY COURSE CONTENT: DISCHOMOTOUS ITEMS..... 28

TABLE 7. BIOLOGY DIF ITEMS BY COURSE CONTENT: DISCHOMOTOUS ITEMS..... 31

TABLE 8. PURE MATHEMATICS DIF ITEMS BY COURSE CONTENT: DISCHOMOTOUS ITEMS..... 35

TABLE 9. POLY-SIB RESULTS: SOCIAL STUDIES 56

TABLE 10. POLY-SIB RESULTS: BIOLOGY 58

TABLE 11. POLY-SIB RESULTS: PURE MATHEMATICS 60

Chapter 1

Introduction

The Grade 12 Diploma Examinations are high school graduation examinations administered in Alberta. The results of these examinations contribute 50% of a student's final blended course mark. The decisions associated with these examinations, such as admission to post-secondary programs or to award scholarships, are high-stakes. Therefore, it is of the utmost importance to ensure the fairness of the examinations administrated to students.

One of the main concerns in ensuring test fairness is the unwanted presence of gender differential item functioning (DIF) attributable to problems with the examination. Gender DIF is present when males and females with the same ability do not have the same probability of correctly answering an item (Hambleton, Swaminathan, & Rogers, 1991). DIF may be attributed to item impact or item bias. An item shows impact if the difference in item performance is caused by knowledge and/or experience that the test is designed to measure. Conversely, an item displays bias if dimensions that are irrelevant to the construct being tested cause the difference in item performance. Item bias could lead to bias in the selection and classification of students (Gokiert & Ricker, 2004). Therefore, it is important that the development of large-scale, high-stakes tests, such as the Diploma Examinations, be as fair as possible for both males and females.

Much research has been conducted to investigate potential sources of gender differential item functioning (DIF) on various standardized tests. Henderson (1999) completed a comprehensive review of studies investigating

gender DIF prior to 1999. Results from these studies showed that gender DIF is related to specific content. Males usually perform better than females on standardized tests of science, mathematics, history, and Social Studies (Doolittle & Cleary, 1987; Doolittle, 1989; Wightman, 1998). In contrast, females usually perform better than males in test of verbal and written abilities (Mazzeo, Schmitt, & Bleistein, 1993; Willingham & Cole, 1997). Henderson summarized several trends in the literature that tried to explain the underlying reasons for these differences:

> Males tend to perform better than females on items that involve proportions, ratios, geometry, graphs, tables, or figures (Burton, 1996, Doolittle & Cleary, 1987; Harris & Carlton, 1993; O'Neill & McPeek, 1993). (Henderson, 1999, p. 33)
>
> Males performed better than females on geometry and mathematics problem-solving items…females performed better on pure mathematics items such as formulas, equations, or theories (O'Neill & McPeek, 1993). (Henderson, 1999, p. 36)
>
> Males tend to perform better than females on items related to science and on items referring to stereotypical male activities. In contrast, females tend to perform better than males on items related to aesthetics and human rights and on items referring to stereotypical female activities (Mazzeo, et al., 1993; O'Neill & McPeek, 1993; Sadker & Sadker, 1994). (Henderson, 1999, p. 33)

In addition to the findings that gender DIF is related to content area, there is also evidence indicating that the presence of DIF may be related to item format:

> Males generally perform better than females on dichotomous items, while females perform better than males on polytomous items like essays (Breland, Danos, Kahn, Kubota, & Bonner, 1994; Pomplun & Sundbye, 1999; Willingham & Cole, 1997). (Henderson, 1999, p. 4)

However, as pointed out by Henderson, findings on gender DIF across content area and item format are still tentative. Results reported in literature on gender

DIF across content area are inconsistent and sometimes contradictory. This may be attributable to the fact that early studies were mainly conducted on several different large standardized tests, where the presence of DIF may be related to the effect of individual item characteristics and/or the types of samples used in the study. Further, DIF studies on tests consisting of both dichotomous and polytomous items were limited, and results reported in literature were not always consistent.

In light of the limitations, Henderson (1999) conducted a comprehensive study in which she examined gender DIF across examinations containing both dichotomous and polytomous items using a common sample. Four diploma examinations representing both the humanities and the sciences were chosen. They were English 30, Social Studies 30, Mathematics 30, and Biology 30. A mixture of dichotomous and polytomous items was used in each Diploma Examination. Within each examination, the test items were described and classified into different content areas according to the test blueprint for each examination. The two samples in Henderson's study included the students who completed all four Diploma Examinations administrated in June 1997 and June 1998.

The proc edure used by Henderson to determine the prevalence of gender DIF across item formats and subject areas was Poly-SIB. Several findings were reported in her dissertation:

1. The DIF prevalence rates for the dichotomous items were similar to those reported in the literature for American high school examinations.

2. Previous findings suggesting that males outperformed females on geometry and mathematical problem solving items and on mathematics items containing graphs, figures, or tables were not found. Similarly, previous findings suggesting that females outperformed males on mathematics items containing formulas, equations, or symbols were not supported.
3. Unlike previous findings, references to stereotypical male or female activities did not consistently favour one group or the other. But most of the items under investigation did not refer to stereotypical activities of either group.
4. No numerical response items were flagged for DIF.
5. A greater number of selection response items favoured males than females, while all the constructed response items identified with DIF favoured females, which is consistent with previous research.

Henderson's dissertation was the first to examine gender DIF using a common sample, where the DIF pattern across item formats and subject areas could be evaluated without the influence of variability across samples. The results of her study provided useful insights for the diploma examination program, especially the last point mentioned above. It verified previous findings that there might be a gender-by-item format interaction where females outperform males on written construct items and males outperform females on multiple choice items regardless of subject areas. However, further studies across samples and testing programs are required to verify this finding.

A main change in the administration procedures of the diploma examinations has been made since Henderson completed her study. The written-response and machine-scored components are now administrated separately for Social Studies, Mathematics, and the Sciences. The intention of this change was to allow students to better demonstrate their ability by focusing on one type of response at a time. However no study has been conducted to examine the impact of the above new policy on gender DIF patterns across subjects and item formats.

Purpose of the Study

Therefore, the primary purpose of the study was to examine the impact of the new administration procedure on the DIF patterns across subjects and item formats for the diploma examinations. A secondary purpose was to examine potential sources of gender DIF using a common sample of examinees across subject areas and item formats.

Definition of Terms

Dichotomous Items

Dichotomous items are items that are scored into two categories: correct or incorrect. In this study, dichotomous items included multiple-choice items and numerical response items.

Differential Item Functioning (DIF)

DIF occurs when examinees from different populations have unequal probabilities of getting an item correct after controlling for differences in the ability being assessed.

Matching Variable

Variable used to match examinees on measured ability so that examinees are comparable on the ability being measured by the test before they are compared.

Focal Group

The primary group of interest to compare against the reference group.

Item Bias

An item displays bias if dimensions that are irrelevant to the construct being tested cause the difference in item performance.

Item Impact

An item shows impact if the difference in item performance is caused by knowledge and/or experience that the test is designed to measure.

Polytomous Items

Polytomous items have more than two score categories. In this study, the polytomous items included written-response items.

Reference Group

A group against which the focal group is compared.

Organization of the Thesis

The balance of this thesis is divided in four chapters. The literature review, presented in Chapter 2, starts with a summary of studies on gender DIF completed since Henderson finished her dissertation in 1999. An overview of the DIF detection method used in this study is then mentioned followed by a description of the old and new administration procedures. The instruments, subject areas, samples, and analysis procedures used in this study are provided in Chapter 3. The

results of analysis are reported and discussed in Chapter 4. A summary of the method and key findings, discussion of the results in terms of previous literature, limitations of the study, conclusions, implications for practice, and recommendations for future research are presented in Chapter 5.

Chapter 2

Literature Review

Chapter 2 is organized in three sections. A summary of findings on gender DIF across content areas and item formats from studies completed after 1999 are described in the first section. In the second section, a review of the DIF detection method used in this study is presented. The last section introduces the new administration procedure implemented by Alberta Education for the Humanities Diploma Examinations in 2003 and the Science and Mathematics Diploma Examinations in 2004.

Gender DIF across Content Areas

Several studies have been conducted to investigate gender DIF due to content and/or item format since Henderson completed her study. A few of them examined the characteristics of mathematics items associated with gender DIF. The results of these studies were inconsistent. Several studies showed that females outperformed males in mathematical algorithms and algebra, while males outperformed females on geometry items (Garner & Engelhard, 1999; Ryan & Chiu 2001; Li, Cohen & Ibarra, 2004; Innabi & Dodeen, 2006). However, studies conducted by Boughton, Gierl, and Khaliq (2000) and Mendes-Barnett and Ercikan (2006) did not find geometry as a source for gender DIF in favor of males. Mendes-Barnett and Ercikan (2006) found that problem-solving items and items containing visuals favored boys on the British Columbia Provincial Principles of Mathematics Examination for Grade 12. Likewise, Ryan and Chiu

(2001) examined the pattern of gender DIF of the items included in the Mathematics Placement Exam administrated in 1996, and found that males outperformed females on items containing figures, graphs or tables.

Turning to other subject areas, Boughton, Dawber, and Hellsten (2001) conducted a differential bundle functioning (DBF) study to examine gender difference on four Social Studies 30 diploma examinations administrated in January and June of 1991 and 1992. Bundles in their study were formed according to the conceptual framework developed by Walter and Young (1997). Five bundles were formed. Four bundles, namely Economics, Politics, History and Control Tactics, were hypothesized to favour males. The remaining bundle, Peace and Internationalism, was hypothesized to favour females. Boughton et al. (2001) found that two bundles, History and Control Tactics, consistently favoured males across the four administrations. The History and Control Tactics bundles were hypothesized to favour males because they mainly contained items referring to stereotypical males activities, such as world wars, atomic weapons deployment or economic strategies. However, while the remaining hypotheses were based on a similar rationale, they were not supported in their study.

Zenisky, Hambleton, and Robin (2003-2004) conducted a DIF study using data from a large-scale state science assessment program across multiple grades and found that multiple-choice (MC) items containing pictures, maps and diagrams favoured males, which is consistent with previous findings (Harris & Carlton, 1993). However, no performance differences between males and females were noted on open-responded items when examinees were asked to diagram the

answer themselves. To sum up, the findings on the latest research provided inconsistent results on gender DIF across content areas: some supported previous findings that were not found in Henderson's study and some agreed with findings reported in Henderson's study. Moreover, although Henderson (1999) pointed out that research controlling sample differences across examinations will help us better understand gender DIF across content areas, these research studies did not involve common samples across different content areas.

Gender DIF across Item Formats

The latest findings on gender DIF across item formats are also inconsistent. Wester and Henriksson (2000) studied the interaction between item format and gender in mathematics using data from the Trends in International Mathematics and Science Study (TIMSS). In their study, the multiple-choice items were changed to an open-ended format. The results showed that gender performance did not change due to the manipulation of item format. Beller and Gafni (2000) also investigated the influence of item formats on gender achievement in mathematics using data from another international assessment, the International Assessment of Educational Progress (IAEP). The results of their study also disagreed with the assertion in previous literature that girls performed relatively better on open-ended items. It should be noted that the open-ended items in these two studies were all short answer items, or a single number or word, instead of extended free response, such as an essay. Moreover, the comparisons between gender was made based on standardized mean scores or item *p* values instead of DIF analyses where comparisons were made using matching ability groups.

In contrast, Garner and Engelhard (1999) conducted a DIF study using data from the mathematics portion of Georgia High School Graduation Test and found that constructed response items tended to favour females. This finding is supported by Bolt (2000) who examined the DIF due to item format using data from the mathematics sections of the Scholastic Assessment Test (SAT). He and found a small but statistically significant effect that disproportionately favoured males when modifying items from their open-ended formats to their multiple-choice formats. Moreover, in their study of potential sources of DIF, Zenisky et al. (2003-2004) found that multiple-choice items tended to favour males while open-responded items tended to favour females. These findings on gender DIF across item formats are consistent with previous research (Breland et al., 1994; Pomplun & Sundbye, 1999; Willingham & Cole, 1997).

Overview of the Poly-SIB procedures

Henderson (1999) compared several of the more popular polytomous DIF detection methods. These methods included the Generalized Mantel Haenzel (GMH), Poly Simultaneous Item Bias Test (Poly-SIB), and Logistic Discriminant Function (LDF). She concluded that Poly-SIB was most liberal and detected the greatest number of DIF items. This was desirable as the purpose of most DIF research was to explore the prevalence of DIF and the characteristics of DIF items. Also, the DIF items detected by Poly-SIB included the most common DIF items identified by all three methods. The Poly-SIB procedure is described below.

Poly-SIB is a general version of SIBTEST that can be used with both dichotomous and polytomous items. The simultaneous item bias test (SIBTEST),

proposed by Shealy and Stout, is used to detect DIF in dichotomous items. The items in a test are divided into two subtests: a "studied" subtest and a "matching" subtest. The studied subtest contains potential DIF items. In a single-item DIF analysis, which is used in this study, the studied subtest only includes one item. The matching subtest usually contains the rest of the items. The total scores examinees obtained from the matching subtest are used to represent examinees' ability, which in the computer program for SIBTEST and Poly-SIB is the sum of scores of the items in the matching subtest (Henderson, 1999). Examinees from the reference group and the focus group are matched on these total scores and grouped into K distinct score level. Examinees within each group are assumed to be equivalent on the ability measured by the test. Then the performance of the examinees is compared across the reference and focal groups on the studied subtest for each of the K groups (Li et al., 1995).

SIBTEST cannot only be used to test DIF hypotheses on an item but also be used to estimate the amount of the DIF. The SIBTEST effect size, $\hat{\beta}_U$, can be interpreted as the amount of DIF for each item. The value of $\hat{\beta}_U$ will be positive if the item favours the reference group and negative if the item favours focal group. $\hat{\beta}_U$ is estimated by

$$\hat{\beta}_U = \sum_{k=0}^{K} p_k (\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk}), (k = 0,1,...K),$$

where

$\hat{\beta}_U$ = estimated amount of DIF – positive values denote DIF against the focal group; negative values denote DIF against the reference group,

p_k = proportion of focal group examinees in subgroup k,

$\bar{Y}^*_{Rk}$ = adjusted mean scores of the studied subtest for the reference group in subgroup k, and

$\bar{Y}^*_{Fk}$ = adjusted mean scores of the studies subtest for the focal group in subgroup k.

SIBTEST adjusts the observed scores on the matching subtest by using "linear regression of true score on observed score from classical test theory with KR 20 calculated as the slope of the regression line for each group" (Li et al., 1995, p.8). The means for examinees in subgroup k are then adjusted using a regression correction procedure described in Shealy and Stout (1993) to ensure the estimated true score is comparable for the examinees in the reference and focal groups on the matching subtest.

The statistic used to test the null hypothesis of no DIF is given by:

$$B_U = \frac{\hat{\beta}_U}{\hat{\sigma}(\hat{\beta}_U)},$$

where $\hat{\sigma}(\hat{\beta}_U)$ is the estimated standard error of $\hat{\beta}_U$:

$$\hat{\sigma}(\hat{\beta}_U) = \left[\sum_{k=0}^{K} \hat{p}_k^{\,2}\left(\frac{1}{N_{Rk}}\hat{\sigma}_R^{\,2}(Y/k,R) + \frac{1}{N_{Fk}}\hat{\sigma}_F^{\,2}(Y/k,F)\right)\right]^{1/2},$$

where

N_{Rk} and N_{Fk} are, respectively, the number of examinees in the reference and focal groups in subgroup *k*, and

$\hat{\sigma}_R^2$ and $\hat{\sigma}_F^2$ are the variances of the studied subtest scores for the reference and focal groups and which are assumed to be equal.

If $\left|\hat{\beta}_U\right| > Z_{1-\frac{\alpha}{2}}$ then the null hypothesis is rejected in favour of the alternative hypothesis that DIF exists for the studied item (Shealy & Stout, 1993).

Roussos and Stout (1996) developed guidelines to interpret $\hat{\beta}_U$ on a single item:

- Negligible or A-level DIF: null hypothesis is rejected and $|\hat{\beta}_U| < 0.059$.
- Moderate or B-level DIF: null hypothesis is rejected and $0.059 \leq |\hat{\beta}_U| < 0.088$.
- Large or C-level DIF: null hypothesis is rejected and $|\hat{\beta}_U| \geq 0.088$.

SIBTEST has been generalized for use with polytomous items (Poly-SIB; Chang, Mazzeo & Roussos, 1996). In Poly-SIB, the *K* in the SIB statistic ($\hat{\beta}_U$) used to estimate DIF is replaced with K_H, where H is the possible score for each item, and K_H is the sum of all item scores (i.e., the maximum possible matching score). Also, KR-20 is replaced by coefficient alpha in performing the regression correction. Poly-SIB uses the same statistic as SIBTEST to test the null

hypothesis of no DIF and the same guidelines are used to interpret effect size $\hat{\beta}_U$ (Henderson, 1999).

Change of Administration Procedure of the Diploma Examinations

As part of an ongoing effort to continually improve Alberta's learning system, Alberta Education changed the manner in which the Diploma Examinations were administered in 2003 for the Social Studies examination and in 2004 for all science and mathematics examinations. Prior to this time, with the exception of English, the dichotomous items (multiple-choice, numeric response when used) and polytomous items (essays, problems) were contained in one examination booklet and were administrated in one sitting. The dichotomous and polytomous items now appear in two booklets. The test booklet containing the polytomous items is administered first, followed by the booklet containing the dichotomous items. In the case of English and Social Studies, approximately one to two weeks separate the two administrations. In the case of Mathematics and Science examinations, the booklet containing the polytomous items is administrated in the morning and the booklet containing the dichotomous items is administered in the afternoon of the same day. It should be noted here that in the case of English, no change was made in administration. The polytomous items have always been administered at least one week prior to the administration of the dichotomous reading comprehension test items.

A preliminary study conducted by Alberta Education in 2004 showed that students performed better in all but one science examination on both the written-response and machine-scored components on the January 2004 diploma

examinations after the new administration procedure was applied for the first time. "These results…show that when students are better able to focus on one type of response, they can better demonstrate their ability," said Minister of Learning, Dr. Lyle Oberg (March 1, 2004). But did this change also have an impact on the DIF patterns for each Diploma Examination? For example, did the DIF prevalence rate increase or decrease under the new administration schedule? Does the new administration procedure benefit one group more than the other? Does the gender-by-item format effect discovered in Henderson's study still exist under the new administration schedule? To date, no study has been conducted to assess the impact of the change in administration procedure on the gender DIF patterns across the item formats and subject areas. Thus, the primary purpose of the proposed study was to evaluate the impact of the change in administration process on gender DIF patterns across item formats and subject areas. A secondary purpose was to examine potential sources for gender differences using new samples.

Chapter 3

Method

In this chapter, the examinations and the student samples included in the study are described first, followed by a description of the analysis procedure.

Diploma Examinations

The Grade 12 Diploma Examinations are high school graduation examinations. Each examination is administered to all students who complete the corresponding Grade 12 course. The Diploma Examinations are high stakes, the results of which contribute 50% to a student's final blended course mark. The examinations are administered four times a year in January, April, June, and August. However, most students write the examinations in either January or June.

To ensure the high quality of the examinations, Alberta Education has a formal review procedure to scrutinize all items during the test development process. One purpose of this review is to examine the test for any possible bias. Two content reviews of each examination are conducted before examinations administration to identify and eliminate gender bias. The first review is conducted before the items to be included in a future examination are field-tested. The items are examined for bias by an internal review committee at Alberta Education. The second review is conducted after the final examination is assembled using the field test items. Each examination is examined for bias by an internal review committee again. Moreover, editors employed by Alberta Education will count the number of references to males and females on the second occasion to ensure the balance between the two groups (Gierl, Khaliq, & Boughton, 1999). However,

while the examinations are carefully examined for potential sources of bias by the staff responsible for the Diploma Examinations, no routine statistical analyses are conducted to assess the possible presence of DIF when the items are field-tested or after they are administered as part of a diploma examination.

Subject Areas

Three of the four subjects studied by Henderson were examined in this study. Mathematics 30 was phased out in 2001 and replaced by Pure Mathematics 30. The other two subjects are Social Studies 30 and Biology 30. The administration procedure for English 30 did not change; therefore, this subject was not included in the current study.

A mixture of dichotomous and polytomous items was used in the three Diploma Examinations considered in the present study. The number of each item type is listed in Table 1.

Table 1
Structure of Examinations

Exam	Number of Items	
	Dichotomous	Polytomous
Social Studies	70	4
Biology	56	2
Mathematics	39	3

2005 was the third year for Social Studies examination and the second year for Pure Mathematics and Biology examinations to be administered using the new procedure. It was presumed that students and teachers were used to the new procedure. The number of dichotomous items was 70 in the Social Studies 30 examination, 56 in Biology, and 39 items in the Pure Mathematics 30

examination. The dichotomous items included in the Social Studies examinations consisted only of multiple-choice items, while the dichotomous items in the Biology and Mathematics examinations included 8 and 6 numerical response items, respectively. The numerical response items involved calculating a numerical answer, recording the understanding of a conceptual idea, and ordering a sequence of listed events. Within each examination, the test items were described and classified into different content areas according to the test blueprint for each examination (Alberta Education web site at www.education.gov.ab.ca).

The number of polytomous items ranged from 2 to 4 items. The Social Studies examinations had the largest number of polytomous items and the Biology examinations had the fewest number. In Social Studies, students were asked to write a complete essay in which they discussed the importance and complexity of an issue and provided evidence to support their position. The essay was marked using four different five-point rating scales, namely "exploration of the issue," "defence of the issue," "quality of the examples," and "quality of the language and expression." Two trained markers independently scored each item. The item score was the average of the scores awarded by the two scorers. In Pure Mathematics, three polytomous items were included on the examination. Students were required to explain mathematics concepts and draw on their own mathematical experiences to solve problems. Each item was scored by one trained marker using a five-point rating scale. There were two polytomous items in Biology, one closed-response question and one open-response question. The closed-response question was a process skill question related to current research

and required students to demonstrate a variety of science process skills. It had several parts. The total number of parts depended on the context. In 2002, the item was presented in five parts, while in 2005, the item had six parts. A closed-response scoring guide was used to mark this question. One marker marked all parts for a total score of 12. Poly-SIB cannot handle scales greater than ten. However, individual scores were available for each part of the question, therefore, DIF analyses were performed for each part. The open-response question contained a problem based on current research that required students to make connections among biological concepts, technology, and social issues. The question was marked using two five-point scoring scales: the science scale and the technology and society scale. Both scales were marked by two trained independent markers. The final score was the average score awarded by the two markers (Alberta Education web site at www.education.gov.ab.ca). DIF analyses are performed for each scale separately.

Samples

The samples for this study included the students who wrote the Social Studies 30, Biology 30, and Pure Mathematics 30 in June 2002 and in June 2005. The June administration was selected because Henderson (1999) used the June administration, thereby allowing comparison of the two sets of results. It was assumed that the students who wrote in June were a more homogeneous group with similar academic and extracurricular interests and completed coursework (Henderson, 1999). There were 243 male students and 406 female students who

wrote all three examinations in June 2002, and 254 male students and 423 female students who wrote all three examinations in June 2005.

Procedure

As mentioned in Chapter 2, Poly-SIB was used in this study to detect DIF items for its advantages over the other polytomous DIF detection methods, and also to allow direct comparison of the results with the results reported by Henderson (1999). Similar to Henderson's study, single-item DIF analyses were conducted for the six data sets (three subjects across two years). Females were the reference group and males were the focal group.

Chapter 4

Results

The purpose of this chapter is to present the results of the DIF analyses and address the research questions raised in the previous chapter, that is, whether the new administration procedure changed the gender DIF pattern and whether the findings on potential sources for gender DIF in this study were consistent with those reported in previous research. The chapter is organized in three major sections. First, performance differences between samples of boys and girls on the set of dichotomous items, the set of polytomous items, and the full set of items are summarized. This is then followed by the overall prevalence of DIF items in each of the years studied. Lastly, the DIF results are discussed in more detail by content area assessed and item format.

Differences between Males and Females on Sets of Items

Differences between the mean scores for males and females are commonly used as reference for differential performance. The means and standard deviations for males and females, *t*-test results, and effect sizes *d* are reported separately for each examination and year in Tables 2 to 4. The effect sizes were interpreted using Cohen's (1988) operational definitions where 0.20 indicated a small effect, 0.50 a medium effect, and 0.80 a large effect. The diploma examinations were administered using the single session procedure in 1997, 1998, and 2002 and the double session procedure in 2005. This is delineated in the tables by the light line.

Social Studies

The mean scores were computed using observed item scores, where each dichotomous item was worth one point and each polytomous item was worth five points. Therefore, the total observed score for the Social Studies examinations was 90: 70 score points for the dichotomous items and 20 score points for the polytomous items. The descriptive statistics for the Social Studies examinations across years are presented in Table 2.

Table 2
Descriptive Statistics for Social Studies Examination across Years

			Total Scores				Dichotomous Items				Polytomous Items			
		n	*M*	*SD*	*t*	*d*	*M*	*SD*	*t*	*d*	*M*	*SD*	*t*	*d*
1997	M	224	62.17	12.59	3.88*	0.35	50.08	10.37	4.96*	0.45	12.10	3.50	-0.89	-0.07
	F	285	57.75	12.88			45.38	10.81			12.36	3.09		
1998	M	183	61.32	13.04	1.45	0.14	49.34	11.27	2.51*	0.24	11.98	3.12	-2.89*	-0.28
	F	243	59.47	13.26			46.61	11.02			12.85	3.05		
2002	M	243	63.69	12.97	4.35*	0.35	49.89	10.63	5.13*	0.42	13.79	3.31	0.03	0.00
	F	406	59.19	12.61			45.40	10.88			13.79	3.00		
2005	M	254	60.81	13.26	4.50*	0.36	46.96	11.64	4.98*	0.41	13.84	2.71	0.81	0.07
	F	413	55.82	14.29			42.17	12.35			13.65	2.97		

Note. * $p < .05$. M=male; F=female.

As shown in Table 2, there was a consistent significant difference between the mean for boys and the mean for girls across the four administrations for the dichotomous items, where boys consistently performed better than girls. The effect sizes ($0.20 \leq |d| < 0.50$) were small, however. No significant mean differences between boys and girls for the polytomous items were found except for the 1998 administration, where girls outperformed boys. Again, the corresponding effect size was small.

Biology

The total observed score for the Biology examination was 78: one point for each of the 56 dichotomous items, 12 points for the first polytomous item, and 10 points for the second polytomous item. As shown in Table 3, no significant mean differences between boys and girls were found for the total scores and the sets of dichotomous item scores across the four years. Girls outperformed boys on the polytomous items in 1998 and 2002, but with a small effect sizes ($0.20 \leq |d| < 0.50$).

Table 3
Descriptive Statistics for Biology Examination across Years

			Total Scores				Dichotomous Items				Polytomous Items			
		n	*M*	*SD*	*t*	*d*	*M*	*SD*	*t*	*d*	*M*	*SD*	*t*	*d*
1997	M	224	53.65	10.75	0.71	0.06	35.02	6.97	0.70	0.06	18.63	4.66	0.62	0.05
	F	285	53.00	9.73			34.62	5.96			18.38	4.46		
1998	M	183	49.71	11.69	-1.83	-0.14	34.97	7.33	-1.04	-0.08	14.74	5.51	-2.68*	-0.21
	F	243	51.39	11.07			35.58	7.12			15.80	4.68		
2002	M	243	53.01	11.11	-0.71	-0.06	38.81	8.04	0.39	0.03	14.21	4.05	-2.86*	-0.23
	F	406	53.63	10.50			38.56	7.82			15.07	3.52		
2005	M	254	52.48	11.44	-0.90	-0.07	38.02	8.86	-0.80	-0.06	14.46	3.47	-0.90	-0.07
	F	413	53.31	11.67			38.59	8.91			14.72	3.71		

Note. * $p < .05$. M=male; F=female.

Pure Mathematics

The Pure Mathematics 30 examination was introduced in 2001. Therefore, data were not available for the 1997 and 1998 administrations. The total observed score for the Mathematics examinations was 54: one point for each of the 39 dichotomous items, and five points for each of three polytomous items. As shown in Table 4, no significant mean differences between boys and girls were found for

the 2002 administration, but boys outperformed girls on the total score and set of dichotomous item score in 2005. Effect sizes ($0.20 \leq |d| < 0.50$) were small, however.

Table 4
Descriptive Statistics for Pure Mathematics Examinations across Years

			Total Scores				Dichotomous Items				Polytomous Items			
		n	*M*	*SD*	*t*	*d*	*M*	*SD*	*t*	*d*	*M*	*SD*	*t*	*d*
2002	M	243	31.97	10.04	-0.29	-0.02	24.11	7.06	0.13	0.01	7.86	3.65	-1.07	-0.09
	F	406	32.20	9.44			24.03	6.63			8.17	3.39		
2005	M	254	32.87	9.19	1.98*	0.16	25.26	7.27	2.44*	0.18	7.61	2.92	0.43	0.03
	F	413	31.35	9.73			23.84	7.41			7.51	2.85		

Note. * $p < .05$. M=male; F=female.

Prevalence of DIF across Subject Area and Item Format

The results of the DIF analyses across examinations and item formats in 1997, 1998, 2002, and 2005 are presented in Table 5 for each of the subject areas considered.

Social Studies. The number of DIF dichotomous items ranged from 8 to 18 when the old administration procedure was used and was 11 in 2005 when the new administration procedure was used. The DIF prevalence rates for dichotomous items were 11%, 26%, and 11% in 1997, 1998 and 2002, respectively, and 16% in 2005. More dichotomous DIF items favoured males, especially in 1997 (6 vs. 2) and 1998 (14 vs. 4). The four polytomous items included in the Social Studies examinations exhibited DIF across all four administrations, and all favoured females.

Table 5
DIF Items across Examinations and Item Formats

		1997		1998		2002		2005	
Subjects	Item format	M	F	M	F	M	F	M	F
Social Studies	D (k=70)	6	2	14	4	5	3	6	5
	P (k=4)	0	4	0	4	0	4	0	4
Biology	D (k=56)	6	6	7	2	2	4	4	7
	P (k=2)	0	0	0	1	0	2	0	0
Pure Mathematics	D (k=39)					5	5	3	4
	P (k=3)					0	1	0	2

Note. D = dichotomous items; P = polytomous items; M = males; F = females;
k = number of items in the analysis

Biology. The number of DIF dichotomous items in Biology ranged from 6 to 12 in the years when the old administration procedure was used, and was 11 in 2005 when the new administration procedure was used. For the dichotomous items, the DIF prevalence rates were 21%, 16%, and 11% in 1997, 1998, and 2002, respectively, and 20% in 2005. An equal number of DIF dichotomous items favoured males and females in 1997, a greater number of DIF items favoured males in 1998 (7 vs. 2), and a greater number of DIF items favoured females in 2002 (4 vs. 2) and 2005 (7 vs. 4). One polytomous item in 1998 and two polytomous items in 2002 were detected with DIF. They all favoured females. No polytomous items displayed DIF in the other years studied.

Pure Mathematics. As mentioned above, the Pure Mathematics Diploma Examination was introduced in 2001; therefore, results were only available for the last two administrations in the present study. For the dichotomous items, the numbers of DIF items were 10 in 2002 and 7 in 2005. The corresponding DIF prevalence rates were 26% and 18%. The DIF items were equally distributed in

2002 and one more item favoured females in 2005. For the polytomous items, one DIF item was found in 2002 and two in 2005, all of which favoured females.

Gender DIF Patterns across Subject Area and Item Format

The DIF results are discussed in more detail for each of the three subject areas in the following three subsections. Within each subject area, the stability of findings between 2002 and 1999 (Henderson, 1999) is first summarized. The changes between 2005 and 2002 are then discussed. Similar to Henderson, no attempt has been made to decide whether the DIF was due to bias, impact, or Type I error.

Social Studies

The 70 dichotomous items included in the Social Studies examinations were multiple-choice items. As shown in Table 6, these items were equally allocated between two content areas – "Political and Economic Systems" and "Global Interaction in the 20th Century" – for each of the four years. The four polytomous items included in the Social Studies examinations were related to one written assignment in which students were asked to discuss the importance and complexity of an issue, decide on their position with respect to the issue, and provide evidence to support their position. As mentioned earlier, the student responses were marked on four different scales, namely "Exploration of the Issue," "Defence of Position," "Quality of Examples," and "Quality of Language and Expression."

Table 6
Social Studies DIF Items by Course Content: Dichotomous Items

Unit	1997		1998		2002		2005	
	M	F	M	F	M	F	M	F
Political & Economic	4 (35)	1 (35)	5 (35)	3 (35)	1 (35)	1 (35)	0 (35)	5 (35)
Global Interaction	2 (35)	1 (35)	9 (35)	1 (35)	4 (35)	2 (35)	6 (35)	0 (35)
Total	6 (70)	2 (70)	14 (70)	4 (70)	5 (70)	3 (70)	6 (70)	5 (70)

Note. The numbers in parentheses are the total numbers of items within the cell classified by course content. The totals are repeated in both the male and female columns. M = males; F = females.

2002 vs. Henderson (1999). In 1997, eight multiple-choice items were identified with DIF, of which six favoured males and two favoured females. Of the six items that favoured males, four were in the content area of "Political & Economic" and two were in "Global Interaction" content area. One item in each area favoured females. In 1998, 18 multiple-choice items displayed DIF, of which 14 favoured males and four favoured females. Of the 14 items that favoured males, five were in "Political & Economic" and nine in "Global Interaction." Of the four items that favoured females, three were in "Political & Economic" and one was in "Global Interaction." In both years, more DIF multiple-choice items favoured males than females.

In 2002, eight multiple-choice items were detected with DIF; the number was the same as that in 1997 and ten less than that in 1998. Of the eight DIF multiple-choice items, five favoured males and three favoured females. Although still a greater number of DIF multiple-choice items were found in favour of males in 2002, the gap between the items that favoured males and those that favoured females was smaller than that in the previous two years. Of the five items favouring males, four were in the content area of "Global Interaction" and one

was in the content area of "Political & Economic." Of the three items favouring females, one was in the content area of "Global Interaction" and two were in the content area of "Political & Economic."

2005 vs. 2002. Eleven multiple-choice items were identified with DIF in 2005, of which six favoured males and five favoured females. In terms of content coverage, while two items displayed DIF in the content area of "Political & Economic" in 2002, with one favouring males and one favouring females, five DIF items were found in 2005, all favouring females. Six items displayed DIF in the content area of "Global Interaction" in both 2002 and 2005. However, four of the DIF items in 2002 favoured males and two favoured females while all six items detected with DIF in 2006 favoured males.

A closer examination of the multiple-choice items that displayed DIF in both years revealed that of the 11 items that favoured males, four contained visual stimuli. Three items required students to examine information in a map and one required students to analyze information in a poster. References to stereotypical male activities were also found in some of the items that favoured males, for example "the use of violence", "military involvement", "defence" and "conflict." These findings are consistent with findings reported in the literature (Burton, 1996; Doolittle & Cleary, 1987; Harris & Carlton, 1993; O'Neill & McPeek, 1993).

Lastly, consistent with previous years, all of the four polytomous items favoured females in both 2002 and 2005.

Biology

The Biology examination consisted of 56 dichotomous items – 48 multiple-choice items and 8 numerical-response questions – and two written-response items. The 56 dichotomous items were classified into six unit topics based on the examination specifications. The total number of items in each of the six topic units ranged from 3 to 18. The number of items in each unit varied slightly across years. As described earlier, the first written-response question was a closed-response question related to a synopsis of current research. It had several parts that assessed students' science process skills. The second written-response question was an open-response question that required students to make connections between biological concepts, technology, and/or social issues. The open-response question was marked on to two scales: a science scale and a technology and society scale, each with a range of 0 to 5.

2002 vs. Henderson (1999). As shown in Table 7, 12 dichotomous items displayed DIF in 1997, of which six favoured males and six favoured females. Most of the DIF items were in the two unit topics with largest number of items. Of the 12 DIF items, five measured "Cell division & Mendelian Genetics," of which two favoured males and three favoured females. Two of the four DIF items in the unit "Nervous & Endocrine System" favoured males and two favoured females. Of the three remaining DIF items, one measured "Molecular Genetics" and favoured females, one measured "Reproductive System & Hormones" and favoured males, and one measured "Population Genetics & Interaction" and favoured males. No DIF items were found in the unit "Differentiation &

Development," which had the smallest number of items. In 1998, nine dichotomous items displayed DIF, of which seven favoured males and two favoured females. Of the nine DIF items, four measured "Molecular Genetics," all of which favoured males; three items measured "Nervous & Endocrine System," of which two favoured males and one favoured females; one item measured "Differentiation & Development" and favoured males; and one measured "Cell division & Mendelian Genetics" and favoured females. No DIF items were found in the two small units "Reproductive Systems & Hormones" and "Population Genetics & Interaction." One polytomous item displayed DIF in 1998, favouring females.

Table 7
Biology DIF Items Organized by Unit Topic: Dichotomous Items

Unit	1997		1998		2002		2005	
	M	F	M	F	M	F	M	F
Nervous & Endocrine System	2 (16)	2 (16)	2 (17)	1 (17)	1 (16)	2 (16)	1 (14)	1 (14)
Reproductive Systems & Hormones	1 (8)	0 (8)	0 (4)	0 (4)	0 (6)	1 (6)	0 (9)	3 (9)
Differentiation & Development	0 (3)	0 (3)	1 (4)	0 (4)	0 (3)	1 (3)	2 (7)	2 (7)
Cell Division & Mendelian Genetics	2 (14)	3 (14)	0 (18)	1 (18)	0 (12)	0 (12)	0 (12)	1 (12)
Molecular Genetics	0 (8)	1 (8)	4 (9)	0 (9)	0 (8)	0 (8)	0 (10)	0 (10)
Population Genetics & Interaction	1 (7)	0 (7)	0 (4)	0 (4)	1 (11)	0 (11)	1 (4)	0 (4)
Total	6(56)	6(56)	7(56)	2(56)	2(56)	4(56)	4(56)	7(56)

Note. The numbers in parentheses are the total numbers of items within the cell classified by course content. The totals are repeated in both the male and female columns. M = males; F = females.

In 2002, only six dichotomous items were detected with DIF, of which two favoured males and four favoured females. The two dichotomous items that favoured males were both numerical response items. One of these items measured "Nervous & Endocrine System" and the other measured "Population Genetics &

Interaction." The four dichotomous items that favoured females were all multiple-choice items, of which two measured "Nervous & Endocrine System", one measured "Reproductive System & Hormones," and one measured "Differentiation & Development." No DIF items were found in the units "Cell Division & Mendelian Genetics" and "Molecular Genetics."

Both polytomous items displayed DIF in favour of females in 2002. The first written-response question was a process skill question related to current research on a genetic disorder. It had five parts. DIF in favour of females was found on parts c and d; no DIF was found on parts a, b, and e. Part c required students to "identify parts of an organ and link the symptoms of the genetic disorder to an abnormal development of this organ." Part d required students to "name hormones and clearly describe the effects each hormone would have on the body" (Alberta Education, 2002, p. 10). The second written response question required students to "discuss various aspects of the use of technology in modern biology" (Alberta Education, 2002, p. 10). The question was marked using a science scale and a technology and society scale. The science scale displayed DIF favouring females.

Compared with the previous two administrations, fewer dichotomous items were detected with DIF in 2002 than in the previous two years. Also, a greater number of DIF items favoured females than males. On the other hand, the polytomous items detected with DIF consistently favoured females.

2005 vs. 2002. Eleven dichotomous items were identified with DIF in 2005. Four of these items favoured males and seven favoured females. Two of the

items favouring males measured "Differentiate & Development," one measured "Nervous & Endocrine Systems," and the fourth measured "Population Genetics & Interaction." The seven dichotomous items that favoured females included five multiple-choice items and two numerical response items. Of the five multiple-choice items favouring females, two measured "Reproductive Systems & Hormones," one measured "Nervous & Endocrine Systems," one measured "Differentiate & Development," and one measured "Cell Division & Mendelian Genetics." Of the two numerical response items favouring females, one measured "Reproductive Systems & Hormones" and the second measured "Differentiate & Development." No DIF items were found in the unit "Molecular Genetics." None of the polytomous items were detected with DIF in 2005.

A greater number of dichotomous items displayed DIF in 2005 than in 2002 (11 versus 6). No item displayed DIF in "Molecular Genetics" in both academic years. Only one item displayed DIF in "Cell Division & Mendelian Genetics" in 2005 and none in 2002. In all other units, DIF items were detected in both academic years. For example, one DIF item was found in the unit "Reproductive Systems & Hormones" in 2002 while three DIF items were found in this unit in 2005. All four items favoured females. Similarly, one DIF item measured "Differentiation and Development" was found in 2002 while four items were found in this unit in 2005. The one item in 2002 favoured females while the four items in 2005 were equally distributed. For the remaining units, the numbers of DIF items identified in the two years were equal or very nearly equal. Two items, one in each year, were detected with DIF in the unit "Population Genetics

& Interaction", both of which favoured males. The unit "Nervous & Endocrine Systems" had three DIF items in 2002 and two DIF items in 2005. One item in each year favoured males and the remaining favoured females.

An examination of descriptions of the items that possessed DIF revealed that females outperformed males on items related to topics that they were familiar with, such as menstrual cycle, breast tumour, and in-vitro fertilization. Four out of nine items that favoured females dealt with such topics. Of the six items that favoured males, two items contained a diagram or histogram.

Pure Mathematics

The Pure Mathematics 30 examination consisted of 39 dichotomous items, including 33 multiple-choice items and six numerical-response questions, and three polytomous items. Based on the examination specifications, each dichotomous item was classified by one of the six unit topics or content domains as shown in Table 8. The total number of items in each of the six topic units ranged from four to nine. The number of items in each unit varied slightly across years. There were three polytomous items in both years, which required students to draw on mathematical experiences to solve problems and explain mathematical concepts.

Table 8

Pure Mathematics DIF Items Organized by Unit: Dichotomous Items

Unit	2002		2005	
	M	F	M	F
Transformations of Functions	0 (6)	1 (6)	1 (6)	0 (6)
Exponents, Logarithms, & Geometrics Series	2 (7)	0 (7)	0 (8)	1 (8)
Trigonometry	1 (8)	1 (8)	0 (9)	0 (9)
Conic Sections	0 (5)	1 (5)	0 (5)	1 (5)
Permutations & Combinations	0 (5)	1 (5)	2 (7)	2 (7)
Statistics and Probability	2 (8)	1 (8)	0 (4)	0 (4)
Total	5(39)	5(39)	3(39)	4(39)

Note. The numbers in parentheses are the total numbers of items within the cell classified by course content. The totals are repeated in both the male and female columns. M = males; F = females.

2002. A total of ten dichotomous items were detected with DIF in 2002. Five multiple-choice items favoured males, four multiple-choice items and one numerical response item favoured females. Of the five multiple-choice items favouring males, two were in the unit of "Exponents, Logarithms, & Geometric Series," two were in "Statistics and Probability," and one was in "Trigonometry." Of the four multiple-choice items and one numerical response item favouring females, one item was found in each of the unit topics except for "Exponents, Logarithms, & Geometric Series."

2005. Seven dichotomous items were identified with DIF in 2005. Of these items, three multiple-choice items favoured males and three multiple-choice items, one numerical response item, and two written response items favoured females. Of the seven dichotomous items detected with DIF, four were in the content areas of "Permutations and Combinations": two favoured females and two favoured males. For the other three DIF items, one favoured males and was in "Transformations of Functions" while the other two items favoured females, with

one in the content area of "Exponents, Logarithms, & Geometric Series" and the other in "Conic Sections."

Compared to 2002, a smaller number of dichotomous items displayed DIF in 2005 (7 versus 10). Two items in "Trigonometry" and three items in "Statistics and Probability" displayed DIF in 2002, while no items in these two areas displayed DIF in 2005. DIF items were found in the remaining units for both academic years. One DIF item was found in the unit "Permutations and Combinations" in 2002 while four DIF items were found in this unit in 2005. The one item in 2002 favoured females while the four items in 2005 were equally distributed between males and females. For the other areas, the numbers of DIF items identified in the two years were equal or nearly equal. Two items, one in each year, were detected with DIF in "Transformations of Functions" and "Conic Sections." The two DIF items in "Conic Sections" all favoured males while the two DIF items in "Transformations of Functions" were equally distributed. "Exponents, Logarithms, & Geometric Series" had two DIF items in 2002, all favouring males, and one DIF item in 2005, favouring females.

In terms of item format, while the three polytomous items identified with DIF, one in 2002 and two in 2005, consistently favoured females, the number of dichotomous items favouring males and females was equal in 2002 and one more item favoring females in 2005. Previous findings that suggested males outperformed females on dichotomous items were not found in these two examinations. However, previous findings that suggested that males tended to perform better than females on items that contained visual stimuli and that

females outperformed males on items that measured pure mathematics were supported by the findings of the present study.

Chapter 5

Summary and Conclusions

The chapter is presented in six sections. The research questions and method used in this study are summarized in the first section. In the second section, key findings are summarized and discussed. Limitations of the study are presented in the third section, followed, respectively, by conclusions, implications for practice, and recommendations for future research.

Summary of Research Questions and Method

The primary purpose of this study was to examine whether the new administration procedure for the Diploma Examinations introduced by Alberta Education in 2003 and 2004 had any impact on the prevalence and patterns of gender DIF patterns across content areas and item formats. Prior to 2003, and with the exception of English 30, the multiple-choice, numeric response (when used), and constructed response items were administered in one testing situation for each examinable subject. With the new procedure, introduced in 2003 for the humanities subjects and in 2004 for the science and mathematics subjects, the constructed response items and the multiple-choice and numeric response items are administered in separate sessions; the constructed response items for the humanities examinations are administered one to two weeks prior to the administration of the multiple-choice and numeric response items, while for mathematics and the sciences the constructed response items are administered in the morning and the multiple-choice and numeric items are administered in the afternoon of the same day. Given this change, the primary purpose of the present

study was to address the following research question: did the change in administration procedure have any impact on the DIF patterns for each Diploma Examination? Three specific questions were addressed: did the DIF prevalence rate increase or decrease under the new administration schedule? Did the new administration procedure benefit one group more than the other? Did the gender-by-item format effect discovered in Henderson (1999), who completed a similar study using the 1997 and 1998 administrations, still exist in the new sample and under the new administration schedule? A second purpose of study was to identify potential sources for gender DIF and compare these sources with those reported by Henderson and in the literature.

To answer these questions, DIF analyses were conducted to investigate the prevalence of gender DIF across item formats for three selected Diploma Examinations, namely, Social Studies 30, Biology 30, and Pure Mathematics 30. The samples used in this study included the students who wrote all three examinations in June 2002, when the Diploma Examinations were last administered using the old procedure, and the students who wrote all three examinations in June 2005, when the examinations were administered in the second or third year using the new procedure. The findings for 2002 were first compared with those reported by Henderson (1999) to examine the stability of DIF results across administrations before the new administration procedure was introduced. The 2002 results were then compared with the results from 2005 to investigate the impact of the new administration procedure on the patterns of gender DIF. Poly-SIB was the DIF detection method used in this study.

Summary of Findings

Prevalence of DIF

As presented in the previous chapter, there was no evidence to indicate that the new administration procedure had a differential impact on the prevalence of gender DIF or on the patterns of gender DIF across subjects and item formats. For the multiple-choice and numeric response items, the DIF prevalence rates across the three examinations considered in the present study ranged from 16% to 20% in 2005, while the DIF prevalence rates varied from 11% to 26% in previous administrations of the same examinations in 1997, 1998, and 2002. In Social Studies, the DIF prevalence rates for the multiple-choice items were 11%, 26%, and 11% in 1997, 1998, and 2002, respectively, and 16% in 2005. More of these items favoured males than females across the four administrations, especially in 1997 and 1998. In Biology, the DIF prevalence rates for multiple-choice and numeric response items were 21%, 16%, and 11% in 1997, 1998, and 2002, respectively, and 20% in 2005. An equal number of these items favoured males and females in 1997, more of them favoured males in 1998, and more of them favoured females in 2002 and 2005. The Pure Mathematics Diploma Examination was introduced in 2001; therefore, the results were only available for 2002 and 2005. The DIF prevalence rates for the multiple-choice and numeric response items were 26% in 2002 and 18% in 2005. The number of these items that favoured females and males was equal in 2002 and nearly equal (3 vs. 4) in 2005.

For the constructed items, the four items included in the Social Studies examination each year were detected with DIF across the four years, and all

favoured females. In Biology, one polytomous item in 1998 and two polytomous items in 2002 were detected with DIF, all of which favoured females. No polytomous item displayed DIF in 1997 and 2005. In Pure Mathematics, one DIF item was found in 2002 and two in 2005, all favouring females.

With respect to content coverage, a greater number of DIF items favoured males than females in the sub content area of "Global Interaction" across the four administrations of the Social Studies examinations. However, while a greater number of DIF items favoured males than females in the content area of "Political & Economic" in 1997 and 1998, an equal number of items favoured males and females in 2002 and all DIF items favoured females in 2005. No consistent patterns of gender DIF across sub content areas were observed in Biology and Pure Mathematics due to the small number of DIF items detected in each sub content area.

Potential Sources of DIF

Nine items, four in 2002 and five in 2005, containing visual content were identified with DIF across the three subject areas, four from Social Studies, three from Biology, and two from Mathematics. Of these nine DIF multiple-choice items, eight favoured males and one favoured females. The visual content in the items that favoured males included graphs, maps, diagrams, histograms, and posters. The only item that favoured females contained a photograph. However, the majority of the items with visual content did not display DIF. While this finding is somewhat consistent with the findings reported in the literature (Burton, 1996, Doolittle & Cleary, 1987; Harris & Carlton, 1993; O'Neill & McPeek,

1993) that suggested items that contained visual stimuli tend to favoured males, Henderson did not find that DIF items with visual stimuli favoured males or, for that matter, females.

Results from Henderson's study also did not support previous research that suggested mathematics items containing formulas, equations or symbols favoured females. In this study, three mathematics items that contained an equation and that displayed DIF, one in 2002 and two in 2005, favoured females. While this finding agrees somewhat with what is reported in the literature (O'Neill & McPeek, 1993), Henderson did not find the same in 1997 or 1998. But like Henderson (1999), the majority of the items containing equations favoured neither males nor females. On the other hand, consistent with Henderson, males did not outperform females on trigonometry items. Of the two trigonometry items detected with DIF in 2002, one favoured males and one favoured females, and no DIF items were found in trigonometry in 2005.

The patterns of DIF on numerical response items included in the Biology and Pure Mathematics did not consistently favour females or males. Four numerical response items exhibited DIF in Biology across the two administrations, two that favoured females and two that favoured males. Two numerical response items were detected with DIF in Mathematics. They both contained an equation and involved calculations to obtain an answer, and both favoured females. These findings are consistent with what was reported by Henderson.

Henderson found that "reference to stereotypical male or female activities" (Henderson, 1999, p. 107) either displayed no DIF or did not

consistently favour males or females, which was different from the finding reported in the literature (Mazzeo, et al., 1993; O'Neill & McPeek, 1993; Sadker & Sadker, 1994). In the present study, reference to stereotypical male or female topics or activities, did seem to be a good variable to explain why some items displayed DIF. For example, in Social Studies, topics involving military, violence, and conflict consistently favoured male students; while in Biology, topics such as menstrual cycle and vitro fertilization tend to favoured female students. But as like Henderson (1999), not many items referred to stereotypical male or female topics or activities were found in the three examinations studied.

Lastly, consistent with Henderson (1999) and previous research (Breland, et al., 1994; Pomplun & Sundbye, 1999), females outperformed males on the constructed response items across the three subjects studied.

Limitations of the Study

The primary purpose of this study was to determine whether the new administration procedure had any differential impact on the gender DIF patterns across content areas and item formats on the Diploma Examinations. Comparison of the DIF results for the same examination administered using the two administration procedures was not possible because new examinations were developed for each administration. However, each Diploma Examination was developed based on the same table of test specifications and using the same item formats in the same numbers. The examinations in each subject area were created to be as similar as possible in terms of content coverage and difficulty. Despite

this, the differences noted above, which were few, may be due to the differences between the items administered in 2002 and 2005.

The DIF analyses were conducted across different examinations using common samples of examinees. Using a common sample to conduct the DIF analyses rules out the influence of sample differences when examining the gender DIF pattern across content areas and item formats. However, while comparisons of gender DIF results with those reported in previous research were made, as pointed out by Henderson (1999), caution should be used when generating the findings beyond the samples of examinees considered in this study and in Henderson's study. Sample differences, combinations of several individual item characteristics, or imperfect systems of measure used in previous research likely account for the contradictory DIF findings in DIF obtained in the previous studies (Bond, 1993; Willingham & Cole, 1997).

Prior to 2003, the Diploma Examinations were released to the schools following their administration. However, beginning with the 2003 administrations, the examinations were made secure to allow equating of the Diploma Examinations within one school year and across school years. Consequently, only item descriptors were available to help explain the gender DIF that occurred in the 2005 examinations considered in the present study. It was not possible to conduct further analyses, such as content analysis. Time prevented conducting interviews and conducting protocol analyses (Ericsson & Simon, 1998) to better identify the source of the DIF found.

Conclusions

Taken together, the results reveal that the prevalence of DIF and the patterns of DIF within content area were similar across the four years examined. The change in the administration schedule whereby the constructed response items are no longer administered in one sitting did not lead to a change in the prevalence and patterns of gender DIF. The item format effect still existed under the new administration procedure where females performed consistently better than males with the same ability on constructed response items across the three subjects studied.

Implications for Practice

The purpose of DIF studies is to ensure fairness and equity in testing. Steps should be taken to conduct DIF analyses when potential examination items are pilot tested and again when the items become operational. At the same time, a policy needs to be developed to determine how to best handle items with DIF that are classified as bias. Bias exists because there is something inherent in the item that favours males or favours females. Inclusion of such items may result in scores that cannot equally validly interpret for both males and females.

Recommendations for Future Research

Inconsistencies were noticed between the results found in this study and the results reported by Henderson for examinations constructed using the same tables of specifications. Further, the results differ from those found in previous research. For example, males did not perform better than females on the dichotomous items in both 2002 and 2005, which was different from the findings

reported by Henderson (1999) and in the literature (Breland et al., 1994; Pomplun & Sundbye, 1999; Willingham & Cole, 1997; Zenisky, et al., 2003-2004). Further, while visual content and "reference to stereotypical male or female activities" (Henderson, 1999, p. 107) seemed to be plausible explanations why some items displayed DIF in this study and in previous research (Burton, 1996, Doolittle & Cleary, 1987; Harris & Carlton, 1993; O'Neill & McPeek, 1993), these findings were not supported by Henderson (1999). These observations have several implications for future research.

First, studies such as the one conducted here and by Henderson (1999) in which common samples were used across content areas and item formats should be conducted to further explore potential sources for gender DIF. The use of common samples helps to rule out the influence of sample differences when studying gender DIF. The study could include other humanities, science, and mathematics examinations to discover what factors or sources of DIF might explain the occurrence of gender DIF.

Second, examination of gender DIF at item level is useful for exploring potential sources for gender DIF, but, as Gierl et al. (2001) suggested, "sources of DIF may be more apparent in patterns across multiple items rather than in performance characteristics associated with single items" (p. 27). Sets of items, or bundles, with common characteristics hypothesized to be associated with gender DIF could be examined to study the effect of differential bundle functioning (DBF). Studying gender DIF in bundles instead of single items could increase the power for detecting group differences. Bundles could be created

using the hypotheses derived from DIF analyses performed at the item level. For example, DBF analyses could be performed for a bundle containing items with visual stimuli to verify whether the bundle favoured males. Also, as mentioned in the previous chapter, it was difficult to summarize the DIF pattern across content areas for some examinations, such as Biology and Pure Mathematics, because of the small number of items that displayed DIF for each content area. DBF analyses could be performed for bundles grouped with items from each content area classified using the test specifications provided by Alberta Education to discover gender DIF pattern across content areas.

Third, Henderson (1999) found that relatively fewer DIF items were detected in Mathematics and Biology examinations as compared to English and Social Studies, and suggested more DIF studies focused on humanities subjects should be done. Since then, Mathematics 30 was phased out, and replaced by Pure Mathematics 30 in 2001. In this study, the DIF prevalence rate for Pure Mathematics was the highest across the two administrations. Therefore, more DIF analyses of the new Pure Mathematics Diploma Examination are warranted.

Fourth, Zumbo (2007) suggested that "testing situation," such as classroom size and administration conditions, has been largely been ignored in previous DIF research as a possible source for DIF. This study, which showed that a change in the administration schedule whereby constructed response items are administered at one time and multiple-choice and numeric response items are administered at a later time (in the week or day) did not change the prevalence or pattern of results, is an example of the studies called for by Zumbo (2007).

Further research into other conditions is needed to test further Zumbo's suggestions.

Lastly, to better understand DIF, think aloud protocols and protocol analyses (Ericsson & Simon, 1998) should be conducted. It has been shown that experts are not able to provide good explanations for why some items display DIF and others do not. Administering a set of items that have been shown to display DIF and items that have been shown not to display DIF in the same content area to a sample of students who are asked to think aloud and then conducting a protocol analyses of the students' responses can lead to better identification of the sources of DIF and whether the DIF is due to bias or impact (Lin & Rogers, 2006). In the case of bias, steps can be taken in the future to avoid such bias.

References

Alberta Education (2002). *Biology 30 Diploma Examination results examiners' report for June 2002*. Edmonton, AB: Author.

Beller, M., & Gafni, N. (2000) Can item format (multiple choice vs. open-ended) *account for gender differences in mathematics achievement? Sex Roles. Vol 42(1-2), 1-21.*

Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using Experimentally designed test items. *Journal of Educational Measurement, 37*, 307-327.

Boughton, K., Dawber, T., & Hellsten, L. (2001, April). *Differential bundle functioning on Social Studies high school certification exams*. Paper Presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Boughton, K. A., Gierl, M. J., & Khaliq, S. N. (2000). *Differential bundle functioning On mathematics and science achievement tests: A small step toward understanding differential performance.* Paper presented at the annual meeting of the Canadian Society for Studies in Education, Edmonton, Alberta, Canada

Breland, H., Danos, D., Kahn, H., Kubota, M., & Bonner, M., (1994). Performance Versus objective testing and gender: An exploratory study of an Advanced Placement History Examination. *Journal of Educational Measurement, 31*, 275-393.

Burton, N. M. (1996). How have changes in the SAT affected woment's math scores? *Educational Measurement: Issues and Practice, 15(4),* 5-9.

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously Scored items: An adaptation of the SIB procedure. *Journal of Educational Measurement, 33,* 333-353.

Cohen, J. (1988). *Statistical power for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Doolittle, A. E. (1989). Gender differences in performance on mathematics achievement items. *Applied Measurement in Education, 2*, 161-177.

Doolittle, A. E., & Cleary, T. A. (1987). Gender differences in performance on mathematics achievement items. *Journal of Educational Measurement, 24,* 157-166.

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity, 5(3),* 178-186.

Garner, M., & Engelhard, G. J. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education. Vol 12(1)* 1999, 29-51.

Gierl, M.J., Bisanz, Jeffrey, Bisanz, G. L., & Boughton, K. A. (2001) Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement, Issues and Practice;* 20 (2), 26-36.

Gierl, M. J., Khaliq, S. N., & Boughton, K. (1999, June). *Gender differential item functioning in mathematics and science: Prevalence and policy implications.* Symposium conducted at the meeting of the Canadian Society for the Study of Education, Sherbrooke, QU.

Gokiert, R. J., & Ricker, K. L. (2004, April). *Gender Differential Item Functioning on WISC-II:Analysis of the Canadian Standardization Sample.* Paper Presented at the annual meeting of the American Educational Research Association, Montreal, QU.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE Publications.

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics Items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6,* 137-151.

Henderson, D. L. (1999). *Investigation of differential item functioning in exit Examinations across item format and subject area.* Unpublished doctor dissertation, University of Alberta, Edmonton, Alberta, Canada.

Innabi, H., & Dodeen, H. (2006). Content Analysis of gender-related differential item functioning TIMSS items in mathematics in Jordan. *School Science and Mathematics, 106(8),* 328-337.

Li, H-H., Nandakumar, R., & Stout, W. (1995, April). *Application of SIB in dealing With issues of DIF in the context of multidimensional data.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items Associated with gender DIF. *International Journal of Testing, 4(2),* 115-136.

Lin, J., & Rogers, W. T. (2006, April). *Validity of the simultaneous approach to the development of equivalent achievement tests in English and French (Stage III).* Paper presented at the annual meeting of National Council on Measurement in Education, San Francisco, CA.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). *Sex-related performance differences on constructed-response and multiple-choice sections of Advanced Placement Examinations* (College Board Report No. 97-2). New York, NY: College entrance Examination Board.

Mendes-Barnett, S., & Ercikan, K. (2006) Examining Sources of Gender DIF in Mathematics Assessments Using a Confirmatory Multidimensional Model Approach. *Applied Measurement in Education, 19(4),* 289-304.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 255-276). Hillsdale, j: Lawrence Erlbaum.

Pomplun, M., & Sundbye, N. (1999). Gender differences in constructed response Reading items. *Applied Measurement in Education, 12,* 95-109.

Roussos, L., & Stout, W. (1996). Simulation studies of the effects of small sample size and studies item parameters on SIB and Mantel-Haenzel type I error performance. *Journal of Educational Measurement, 33,* 215-230.

Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14(1)*, 73-90.

Sadker, M., & Sadker, D. (1994). *Failing at fairness. How our schools cheat girls.* Toronto, ON: Simon & Schuster.

Shealy, R., & Stout, W. F. (1993). An item response theory model for test bias. In P.W. Holland & H. Wainer (eds.), *Differential Item Functioning* (pp.197-239). Hillsdale, NJ: Lawrence Erlbaum.

Walter, C., & Young, B. (1997). Gender bias in Alberta social studies 30 examinations: Cause and effect. *Canadian Social Studies, 31,* 83-89.

Wester, A., & Henriksson, W. (2000). The interaction between item format and Gender differences in mathematics performance based on TIMSS data. *Studies in Educational Evaluation 26*, 79-90.

Wightman, L. F. (1998). An examination of sex differences in LSAT scores from The perspective of social consequences. *Applied Measurement in Education, 11*, 255-277.

Willingham, W.W., & Cole, N. S. (1997). Fairness issues in test design and use. In W.W.Willingham & N. S. Cole (eds.), *Gender and Fair Assessment* (pp. 227 – 346). Hillsdale, NJ: Lawrence Erlbaum.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003-2004). DIF detection and Interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment, 9(1&2)*, 61-78.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly, 4(2)*, 223-233.

Appendix A

Poly-SIBTEST Results for Each Examination

Table 9

Poly-SIB Results: Social Studies 30

	2002			2005		
Item	Beta		DIF	Beta		DIF
MC1	0.041		1	-0.006		1
MC2	-0.065		1	-0.066		1
MC3	0.066		1	-0.075		1
MC4	0.018		1	0.030		1
MC5	-0.078		1	0.044		1
MC6	-0.049		1	-0.014		1
MC7	0.059		1	0.030		1
MC8	0.053		1	0.004		1
MC9	0.008		1	0.001		1
MC10	0.051		1	0.010		1
MC11	0.025		1	0.007		1
MC12	0.016		1	-0.026		1
MC13	0.001		1	0.104	*	3
MC14	0.018		1	-0.031		1
MC15	0.101	*	3	0.010		1
MC16	0.020		1	-0.052		1
MC17	-0.001		1	0.001		1
MC18	-0.032		1	0.005		1
MC19	0.014		1	0.071		1
MC20	-0.141	*	3	-0.046		1
MC21	-0.038		1	-0.004		1
MC22	0.083		1	0.041		1
MC23	0.026		1	0.057		1
MC24	0.045		1	0.026		1
MC25	-0.016		1	0.016		1
MC26	-0.060		1	0.091	*	2
MC27	-0.018		1	0.039		1
MC28	-0.030		1	0.004		1
MC29	-0.059		1	-0.023		1
MC30	0.031		1	0.030		1
MC31	-0.035		1	0.091	*	2
MC32	-0.033		1	-0.008		1
MC33	0.024		1	0.179	*	3
MC34	0.046		1	0.040		1
MC35	0.030		1	0.079	*	2
MC36	-0.109	*	3	0.001		1
MC37	0.077	*	2	0.003		1
MC38	0.040		1	-0.079		1
MC39	0.037		1	-0.078		1

Note. * = $p < .05$; 1 = negligible or no DIF; 2 = moderate DIF; 3 = large DIF.

	2002			2005		
Item	Beta		DIF	Beta		DIF
MC40	0.006		1	0.015		1
MC41	-0.021		1	-0.035		1
MC42	-0.028		1	0.002		1
MC43	0.051		1	-0.154	*	3
MC44	-0.042		1	-0.160	*	3
MC45	-0.050		1	-0.073		1
MC46	0.047		1	-0.113	*	3
MC47	0.084	*	2	-0.015		1
MC48	0.077		1	-0.013		1
MC49	0.032		1	-0.022		1
MC50	0.012		1	-0.082		1
MC51	-0.030		1	-0.048		1
MC52	-0.034		1	-0.045		1
MC53	-0.104	*	3	-0.045		1
MC54	-0.098	*	3	0.025		1
MC55	-0.015		1	0.027		1
MC56	-0.057		1	0.014		1
MC57	-0.005		1	-0.001		1
MC58	0.072		1	-0.059		1
MC59	-0.022		1	-0.052		1
MC60	0.022		1	-0.112	*	3
MC61	0.007		1	-0.013		1
MC62	0.047		1	0.012		1
MC63	-0.020		1	-0.025		1
MC64	0.005		1	-0.005		1
MC65	-0.173	*	3	-0.073		1
MC66	-0.057		1	-0.034		1
MC67	0.020		1	-0.089	*	2
MC68	0.002		1	-0.039		1
MC69	0.035		1	-0.088	*	2
MC70	-0.015		1	0.017		1
WR1	0.236	*	3	0.145	*	3
WR2	0.242	*	3	0.130	*	3
WR3	0.251	*	3	0.121	*	3
WR4	0.233	*	3	0.131	*	3

Note. * = $p < .05$; 1 = negligible or no DIF; 2 = moderate DIF; 3 = large DIF.

Table 10

Poly-SIB Results: Biology 30

	2002			2005		
Item	Beta		DIF	Beta		DIF
MC1	0.076	*	2	-0.040		1
MC2	0.073		1	0.082	*	2
MC3	0.008		1	-0.023		1
MC4	-0.058		1	-0.059		1
MC5	-0.061		1	-0.062		1
MC6	-0.063		1	-0.065		1
MC7	-0.060		1	0.039		1
MC8	-0.011		1	-0.051		1
MC9	-0.020		1	-0.079	*	2
MC10	-0.029		1	-0.008		1
MC11	0.010		1	0.012		1
MC12	-0.011		1	0.016		1
MC13	0.004		1	0.023		1
MC14	-0.002		1	0.048		1
MC15	-0.036		1	0.010		1
MC16	0.032		1	-0.015		1
MC17	0.089	*	3	-0.002		1
MC18	0.072		1	-0.031		1
MC19	0.005		1	0.123	*	3
MC20	0.101	*	3	0.063		1
MC21	-0.012		1	0.141	*	3
MC22	-0.004		1	-0.001		1
MC23	0.004		1	-0.015		1
MC24	-0.055		1	0.036		1
MC25	0.003		1	0.102	*	3
MC26	0.017		1	-0.130	*	3
MC27	-0.017		1	0.005		1
MC28	-0.066		1	0.070	*	2
MC29	-0.003		1	0.009		1
MC30	-0.03		1	-0.074	*	2
MC31	0.073		1	-0.055		1
MC32	0.027		1	0.030		1
MC33	-0.022		1	0.001		1
MC34	-0.013		1	-0.054		1
MC35	-0.039		1	-0.016		1
MC36	-0.008		1	-0.018		1
MC37	-0.052		1	0.049		1
MC38	-0.062		1	0.023		1
MC39	0.076		1	0.022		1

Note. * = $p < .05$; 1 = negligible or no DIF; 2 = moderate DIF; 3 = large DIF.

	2002			2005		
Item	Beta		DIF	Beta		DIF
MC40	-0.020		1	-0.005		1
MC41	0.064	*	2	-0.010		1
MC42	-0.001		1	0.037		1
MC43	-0.056		1	-0.006		1
MC44	-0.047		1	-0.029		1
MC45	-0.028		1	-0.043		1
MC46	-0.018		1	-0.032		1
MC47	0.049		1	-0.026		1
MC48	-0.011		1	-0.052		1
NR1	-0.048		1	0.015		1
NR2	-0.136	*	3	0.068		1
NR3	0.000		1	0.061	*	2
NR4	0.056		1	0.050		1
NR5	0.079		1	0.086	*	2
NR6	0.025		1	-0.073		1
NR7	0.007		1	-0.025		1
NR8	-0.120	*	3	-0.057		1
WR1	-0.075		1	0.062		1
WR2	0.072		1	0.079		1
WR3	0.175	*	3	0.045		1
WR4	0.205	*	3	0.065		1
WR5	-0.060		1	-0.032		1
WR6	0.317	*	3	-0.037		1
WR7	0.108		1	0.120		1
WR8				-0.148		1

Note. * = $p < .05$; 1 = negligible or no DIF; 2 = moderate DIF; 3 = large DIF.

Table 11
Poly-SIB Results: Pure Mathematics 30

	2002			2005		
Item	Beta		DIF	Beta		DIF
MC1	0.062		1	0.043		1
MC2	-0.057		1	0.015		1
MC3	0.044		1	0.026		1
MC4	-0.046		1	0.039		1
MC5	0.072	*	2	-0.085	*	2
MC6	-0.187	*	3	-0.059		1
MC7	-0.023		1	0.004		1
MC8	0.011		1	-0.043		1
MC9	-0.117	*	3	-0.002		1
MC10	-0.007		1	-0.048		1
MC11	-0.017		1	-0.047		1
MC12	-0.041		1	-0.037		1
MC13	0.023		1	-0.058		1
MC14	0.050		1	-0.064		1
MC15	-0.009		1	-0.066		1
MC16	0.176	*	3	0.027		1
MC17	-0.099	*	2	-0.050		1
MC18	-0.033		1	-0.030		1
MC19	-0.059		1	-0.023		1
MC20	0.045		1	-0.015		1
MC21	0.064		1	-0.012		1
MC22	-0.057		1	0.033		1
MC23	-0.042		1	0.084	*	2
MC24	-0.050		1	-0.034		1
MC25	-0.012		1	0.091	*	2
MC26	0.077	*	2	0.052		1
MC27	0.095	*	2	0.089	*	2
MC28	-0.115	*	3	0.068		1
MC29	-0.036		1	-0.100	*	3
MC30	-0.126	*	3	-0.068	*	2
MC31	-0.055		1	-0.003		1
MC32	0.027		1	-0.046		1
MC33	0.022		1	0.007		1
NR1	0.008		1	-0.013		1
NR2	0.069		1	0.074	*	2
NR3	0.069		1	-0.017		1
NR4	0.110	*	3	0.037		1
NR5	0.019		1	-0.058		1
NR6	0.011		1	-0.024		1
WR1	0.096		1	-0.162		1
WR2	-0.082		1	0.230	*	3
WR3	0.333	*	3	0.254	*	3